Introducing Language Engineering Tools to Support Information Processing in Healthcare Telematics

Werner Ceusters, Maarten Laga
Language and Computing nv, Hazenakkerstraat 20A, B-9520 Zonnegem, Belgium
phone: +32 53 62 95 45 - email: werner@landc.be - http://www.landc.be/

Abstract. Computer applications have become an essential part in the delivery of care and it is expected that their impact will even grow more in the future. Hence the need for more advanced ways of communication guided by natural language. In this paper, we present an introductory overview of the possibilities of medical natural language processing and the enabling technologies that are available today.

1. Introduction

Faithful recording of patient data can only be achieved by using natural language. This was already stated in the early eighties by Wiederhold who claimed that *the description of biological variability requires the flexibility of natural language and it is generally desirable not to interfere with the traditional manner of medical recording* [1]. At the other hand, it is evenly true that without proper mechanisms in place, free natural language registrations are impossible to be understood by machines, if not to say, quite often also by colleagues. Very often, medical statements are written down in a context that is obvious at the time of registration, but that is difficult to reconstruct later on by third parties, or even by the original source. Also, in order to allow a computer to process healthcare data further, the data must be available in a coded and structured format. Making that happen in a transparent way for healthcare specialists, is the ultimate goal if not even the “raison d’être” of natural language understanding applications in healthcare.

Natural language processing systems are already looked at since the sixties, though mostly only in academic environments. Now it is recognised by major technology consultants in the healthcare domain as an emerging technology with great prospects for the near future. Real applications start to become available, and once the current problems related to continuous speech recognition will be solved, a massive penetration of natural language understanding applications will undoubtedly occur.

In this paper, we give an overview of the many faces of linguistic engineering applications, focusing on their impact on healthcare informatics and telematics today and tomorrow. It is not the idea to give a detailed course in theoretical computational linguistics. In line with the new ideas on successful management in business, we prefer to pay attention to solutions, while not loosing time by focusing too much on the problems [2].

2. Babylon revisited

The development of a machine that understands a human being has been a great dream ever since the beginning of computers. Proof of that are the numerous science-fiction story’s in which a computer is addressed in ordinary human speech, i.e. natural language, upon which the machine promptly answers with a metallic sounding voice. The obstacles between that dream and today’s reality are still enormous, but the light is beginning to shine in certain specialised domains.

Natural language processing applications come in many flavours. At the heart of the technology is a specific discipline of science called *computational linguistics*, aiming to develop computational models of language that explain how language works in human beings, and how this insight can be used to allow computers to work with language. If the focus is more on the development of practical applications rather than on theoretical studies, the term *linguistic engineering* is preferred. That is what we are dealing with in this paper.

As with many disciplines, sub-branches of linguistic engineering emerged very quickly. A first major division is to be recognised between *language processing* and *speech processing*. The basic aim of *speech processing* is to turn the sound wave generated by a speaking human being into a digitally represented text, f.i. by using the ASCI set of characters. The result can be used in word processors or printed on paper. The computer processing the speech signal has however no understanding of the meaning of what has been said, nor is the resulting text by any means a representation that is immediately understood by the machine. *Language processing* at the other hand starts with the verbal representation of - say - an ASCI text, and uses this format to do some further useful processing.

A second major subdivision that cuts orthogonally through the previous one, is whether or not understanding of speech or language is at stake. It is possible to do many tricks with language - and even to build very useful applications by doing so - without a need for true understanding of spoken or written texts. Many information retrieval packages operate in this way by doing string searches, some basic stemming procedures - such as transforming conjugated verbs or plurals to their base form - and counting words, with fairly adequate success. Also the *command & control* paradigm where a computer user can dictate commands to a computer instead of using a mouse, belongs to this class. For this kind of applications, the general terms natural language processing, versus speech processing apply, whereas if true understanding is achieved, the term *natural language understanding* is preferred.

A third division has to do with the direction of processing. While generally with natural language understanding, *natural language analysis* is understood (going from a text to its meaning), the opposite (going from a meaning representation to a text) is called *natural language generation*. For speech applications, the terms *speech-to-text* or *text-to-speech* are often used. Be aware that also here the understanding issue cuts orthogonally through the applications. It is perfectly possible to have text-to-speech applications that do not understand what is being said. Also specific paradigms of *machine translation* work quite well without understanding.

Natural Language Understanding is being considered as one of the most complex problems in artificial intelligence. Up to now, a computer is not yet capable of really understanding the true meaning of ordinary human language. The necessary background knowledge is so extensive and complex that even today's description-possibility's are unable to describe everything. Food for thought is the idea that a human child needs at least six years to adopt a language and that even today's supercomputers don't posses even a fraction of the comprehension capability's of the human brain.

However ! Under certain specific circumstances it is possible to have a computer understand natural language. Medical language, as a sub-language of ordinary human language, is a field that complies in an excellent way with the 'specific circumstances' required: a closed world with restricted domains and disciplines easily separated from each other, a relatively uniform terminology, and the availability of numerous descriptions (textbooks, classifications, …). Because the principles of understanding natural language in the world of medicine could have immediate and huge advantages, the conception of systems to make a machine understand medical language has been a field of research for nearly 20 years. The results of this research are now becoming available as *medical language technology*, and this is the true topic of this paper.

3. Natural language understanding applications for healthcare telematics

There are numerous applications for which medical language technology may pay off. Quite a bit of those have an immediate added value in the present and future clinical-care organisations, most often as enabling tools in the field of traditional telematics. Medical Language Technology is the new engine that will provide the power to stimulate the next generation of medical software applications. Table 1 summarises the possibilities. Some are discussed more deeply in the following paragraphs.

Table 1: Use of NLU in the healthcare domain

- semi- and full-automatic ICD-registration and coding based on full-text-reports.
- medical terminology-management on all levels (departmental, hospital, HMO, National)
- natural-language data-entry-facilities for EPR-systems
- tools for building, selection and evaluation of clinical guidelines on all levels (departmental, hospital, HMO, national)
- automatic translation of medical files into a multiple range of languages (for telematic or telemedicine-purposes)

- automatic conversion of medical files into different classifications and mapping between classifications (ICD9, ICD10, Snomed, CPT4, UMLS, Mesh, Read, ICPC, ICNP, CISP, …)
- tools for medical-data-cleaning and uniformisation for datawarehouses
- tools for full-text-retrieval and semantic searches
- access tools for internal and external knowledge-bases

3.1 Automatic encoding

To overcome the problems related to the use of natural language in communication and clinical registration, coding and classification systems have been introduced as interlingua. Systems such as ICD, Snomed International, ICPC, CPT and many others are now widely used to register medical findings, diagnoses or procedures. Similarly, terminological systems such as NIC, NANDA, ICNP and others are proposed to be used as interlingua in a nursing environment.

Coding patient data means that a physician (or professional encoder) has to describe the patient data by means of codes that are a kind of placeholders for the concepts available in systems such as ICD. The requirements to be met in order to perform the coding task adequately are [3] : 1) a perfect understanding of the meaning of the patient data (the source concepts), 2) a perfect understanding of the meaning of the concepts available in the concept system (the target concepts), 3) at least a certain level of similarity and coherence between the source concepts and target concepts, 4) facilities to search the concept system for the target concept(s) that match(es) a given source concept as closely as possible.

It is common knowledge that coding performed by humans is of rather low quality, both in terms of recall/precision, inter-rater variability, and even reproducibility by the same team. Natural language understanding tools can improve coding quality dramatically.

3.2 Medical terminology management

Coded data are the most convenient way for computers to turn data into information. This is the main reason for the success of coding and classification systems. Hélas, the one omni-potent classification system that fulfils the needs of all doctors, nurses, hospital managers, governments, librarians and international organisations, has yet to be developed. Allow us to speak freely: we are quite convinced it never will be built ! There always will be a need for local variations, for additional dimensions, for greater detail, etc. And as long as a variety of systems continues to be available, the need for integration, mappings and translations will also continue to exist.

That is why people working in the domain of medical natural language processing invest in the development of tools that allow them to work with various classifications, without however becoming too much dependent on them. Assisted by such language analysis tools, mappings can be created from local systems to any other, while guaranteeing that they will remain compatible with future and previous versions. By doing so, users can be sure that their precious data don’t become worthless once a new version of an official classification system becomes available.

3.3 Natural language data entry

Continuous speech recognition software will soon become available at a level of quality that is acceptable to be used in routine medical practice. Discontinuous speech applications are on the market since many years but cannot be said to be a big success. Speaking discontinuously, i.e. pausing after each word or word cluster, is not really practical. Also “command and control” speech applications where - if we may say so - not just the keyboard is replaced by a microphone, but also the mouse movements are to be guided by the voice (“go to medication”, “enter 3 tablets of Aspirin”, “go back”, …) are only useful in some uncommon situations where it is impossible to use the hands to operate a mouse, light pen, keyboard, or whatever other “conventional” input device.

The availability of continuous speech recognition software will have as consequence that the structured data entry of today will disappear gradually, probably even completely in a not so far future. This requires for powerful full text understanding systems that can capture the true semantics of what is said by the user. For specific domains (radiology, pneumology, …), such “text-to-meaning” applications are already available, and this in various languages. Interest in such systems (e.g. [4]) is constantly growing thanks to XML, a format that is perfectly suited to capture the recursively embedded meaning-representations resulting from free text analysis.

3.4 Clinical trials and practice guidelines

Language understanding services are needed when free text entries (whether being full text or short phrases) entered in a certain context, are to be used for other purposes. A typical example is matching patient selection criteria for clinical trials. It is not easy for a physician seeing patients on a routinely basis to bear in mind constantly what clinical trials are running in his department, and what criteria must be met by a patient to enter a trial. It is not feasible to run over the inclusion criteria for each single patient during an encounter. It is more sensible to have a software "watchdog" that constantly monitors the data entered by a physician, and that produces an alert when specific criteria are met. If data are entered in free text, this means that such a watchdog must have enough language understanding power to identify "numbness in left lower leg since last week" as satisfying an inclusion criterion such as "sensory disorders of the limbs lasting for more than 24 hours".

The same goes for checking whether or not practice guidelines are followed when registering patient data, or to generate other alerts upon specific criteria.

3.5 Intelligent querying and information retrieval

Many electronic patient record systems keep collections of text documents (discharge summaries, referral letters, surgery reports, …) related to individual patients. Documents in these "result servers" are only accessible through general indicators such as the original source, the kind of document or the creation data. Searching documents on the basis of their content is seldom possible, or only by means of string search or some crude pattern matching mechanisms with jokers. Natural language understanding techniques can add a lot of functionalities to these primitive mechanisms.

Searches could be improved by using a thesaurus. A basic problem is however where to get one that is suitable for your needs. For medical bibliographic retrieval, one could use the well known MESH thesaurus from the National Library of Medicine. But this thesaurus is largely insufficient to be used in clinical practice. With the proper natural language processing tools, special purpose thesauri that respond to local demands can be built.

Having a thesaurus is not enough. The next step is to attach thesaurus entries to the documents. This is the problem of indexing. Traditionally, indexing is done manually. Professional indexers read a document, and assign the relevant thesaurus entries to it. Natural language understanding software is able to automate this process partly or even completely. The result is an electronic index that gives you fast access to documents on the basis of their "conceptual content" and not limited to the occurrence of specific words.

However, this is not the end of the story. Using a thesaurus to index properly a collection of documents, guarantees that you will find all (and not more) the documents that you need, provided that you know perfectly the terminology used in the thesaurus. To overcome this restriction, "query enhancement" techniques can be used to match a user's query to one or more relevant thesaurus entries. This requires the use of a semantic network.

4. Enabling technologies for natural language understanding

In order to achieve their mission, medical language engineering companies have to develop or acquire various technologies that are indispensable in the process of representing medical natural language in a format understandable by machines. These technologies are used in-house to expand the knowledge resources they are gradually building up, or are embedded in linguistic middleware applications developed for their clients. We refer to these resources and technologies together as "lingware" because contrary to traditional informatics tools, they are specifically designed for linguistic processing, and software and knowledge bases are tightly interconnected.

4.1 Automatic knowledge extractors

Linguistic engineering is a vary labour intensive activity. Hence there is a need for tools that can be used to automatically extract knowledge from text documents. Whether they are in English, Dutch, French or whatever other European language, the vast majority of typical expressions contained in documents pertaining to a specific domain should be extracted on the fly. In addition, semantic relationships between the content words of the documents (i.e. those words pertaining to the domain) are to be made explicit.

Language & Computing nv for instance uses such extractors in-house to continuously update the medical linguistic knowledge base on the basis of new documents coming from electronic textbooks, papers, medical records or discharge summaries. They can also be used outside the laboratory in end user applications for real-time text analysis.

4.2 Machine readable multilingual medical lexicons

Dictionaries are usually large books intended to be used by humans to look up the meaning of unknown words. Most electronic dictionaries currently available differ only from paper dictionaries in their being published on a digital medium. The major advantage is that they can be used from within the most popular word processors without the need for retyping. But their audience consist still of human readers…

For medical natural language understanding purposes, dictionaries have to be fundamentally different in nature: they are primarily intended to be used by machines ! Such dictionaries can be used by some of the knowledge extraction software to represent the meaning of full text documents. But they also can be integrated in third party systems for information retrieval, spell checking, automatic translation, etc.

4.3 Formal medical concept system

What makes the bespoken lexicons even more valuable, is that entries are not related directly to each other, but through a language independent formal medical concept system. The concept system that is developed within our laboratory, is a true model of medicine (at least the parts of it that currently are covered), and is implemented in a powerful semantic network with formal controls. An important characteristic of this system is the clean separation of different kind of relationships that hold between the concepts. Many systems use just “medicine-ignorant” relationships such as “broader than”, “synonym of” or “part of”. Others do have medical relationships specified, but mixed with the taxonomic “is-a” relationship, eg “penicillin is-a drug” and “penicillin is-a reason for an allergic reaction”. These systems are capable of some reasoning, but are very rapidly caught in wrong path flows.

Our medical concept system can be used for various purposes. In home, we use it to improve the recognition capabilities of our language analysers, or to link texts to external databases. It is this concept system that allows us to link various coding- and classification systems to each other, or to integrate the local detailed classification systems that are developed in specialised centres. And why not, also to the concept systems developed by third parties, provided they are “open” …

4.4 Medico-linguistic ontology

We are sure not to be unique in the development of a medical concept system. We are nevertheless happy to belong to a fairly small group that does this in a rigorous and formal way. And we are proud that according to our knowledge, we are probably the only company in which searching for a “language independent” system, did not led us to ignore language, a mistake quite often committed by people working in that field. If the concept system is solely intended to be used as a knowledge base for internal processing, without any communication being needed in natural language, then there are some arguments for such an approach. In the opposite case, it will definitely lead to unsatisfactory behaviour.

Why is it then so important to keep language in mind when developing a language independent concept system (or semantic network) for healthcare ? Simply because the taxonomic and other relationships that hold between two medical concepts are not necessarily reflected in the way people speak about them. Formally, a concept system can have defined the part-of relationship being transitive over long chains, such that a fingernail is part-of a finger, which is-part of the hand, which is-part of the lower-arm, which is part-of the arm, hence a fingernail is-part of the arm. It sounds however very odd if people would say “I lost two fingernails of my left arm”.

Here is another example: it is appropriate to have in a formal medical concept system the notions of “filling” and “injecting” being specific kinds of “resource-installing-procedures”. But for instance in English, the allowed syntactic-semantic constructions quite differ. In a sentence such as “the X is filled”, the X almost stands for that item into which something is installed. However, in a sentence such as “the X is injected”, X can stand both for the substance being injected as for the item into which a substance is injected. Just look for your self: “the arm is injected”, and “the penicillin is injected”.

Our approach is to keep the concept system separate from any linguistic knowledge. But in addition, for each specific language (English, Dutch, French, …) a linguistic ontology is maintained, capturing the relationships between the grammars of these languages, and the language independent concept system [5, 6, 7, 8].

5. Conclusion

Medical information systems are sufficiently large and varied such that no one vendor can expect to provide all of the systems needed in even a single hospital, let alone for the health service as a whole. Many of these varied systems would benefit from natural language interfaces and some, such as automatic linkage to abstracts of the literature, are even impractical without it. Generic multilingual solutions are required if the range of services to be built is to meet the demand. Furthermore, it is essential that the natural language processing components share the underlying concept structure used by the various applications.

Electronic patient record systems are no exception to this. A wealth of knowledge is needed to enter information in those systems consistently and to use the information afterwards for various purposes. Provided that a highly acceptable system can be designed in a specific environment, then developers surely will want to make it available to other users. Whilst much re-use of system components is feasible within a given market segment, there are significant costs associated with the 'localisation' of systems to the needs of other markets. Perhaps the most important of these costs is the localisation to the linguistic needs of each national market in Europe.

Medicine is a descriptive, language intensive activity, and the costs of developing, and perhaps more importantly maintaining, the linguistic resources needed to localise clinical systems are clearly high. Any practical approach to the management and exploitation of linguistic resources in large scale clinical information systems must be based on common methods and internal representations for linguistic information. This information must be reusable across a wide range of systems and local variants of those systems, and the cost of maintaining that information must be separable from those of maintaining the rest of the system.

6. References

[1] Wiederhold G. Databases in healthcare. Stanford University, Computer Science Department, Report No. STAN-CS-80-790, 1980.

[2] Drucker P, Dryson E, Handy Ch, Saffo P, Senge P, Looking ahead: Implications of the present. In: Harvard Business Review, sept-oct 1997.

[3] Ceusters W, Lovis C, Rector A, Baud R. Natural language processing tools for the computerised patient record: present and future. In P. Waegemann (ed.) *Toward an Electronic Health Record Europe '96 Proceedings*, 294-300, 1996.

[4] Ceusters W, Spyns P, De Moor G, Martin W (eds.) Syntactic-Semantic Tagging of Medical Texts: the Multi-TALE Project. Studies in Health Technologies and Informatics, IOS Press Amsterdam, 1998.

[5] Ceusters W, Deville G, Buekens Ph. *The Chimera of Purpose- and Language Independent Concept Systems in Health Care*. In Barahona P, Veloso M, Bryant J (eds.) Proceedings of the XIIth International Congress of EFMI, 208-212, 1994.

[6] Deville G, Ceusters W. *A multi-dimensional view on natural language modelling in medicine: identifying key-features for successful applications*. In: Proceedings of IMIA WG6, Vevey 29/05 - 01/06/94, Switzerland.

[7] Ceusters W, Deville G, Streiter O, Herbigniaux E, Devlies J. *A Computational Linguistic Approach to Semantic Modelling in Medicine*. In: Beckers WPA, ten Hoopen AJ (eds) Proceedings of MIC'94, Velthoven, The Netherlands, 25-26/11/94, 311-319, 1994.

[8] Ceusters W, Buekens F, De Moor G, Waagmeester A. *The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition.* Proceedings of IMIA WG6, Jacksonville, Florida, 19/01/97.